Estimation of structural similarity of membrane proteins by hydropathy profile alignment

Lolkema, Juke S.; Slotboom, Dirk-Jan

Published in:
Molecular Membrane Biology

DOI:
10.3109/09687689809027516

Publication date:
1998

Molecular Membrane Biology, 1998, **15**, 33–42

Estimation of structural similarity of membrane proteins by hydropathy profile alignment

Juke S. Lolkema* and Dirk-Jan Slotboom

Department of Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands

Summary

Many membrane proteins consist of bundles of α-helices that are reflected in typical hydropathy profiles of the amino acid sequences. The profiles provide a link between the amino acid sequence of the polypeptide chain and its folding and are much better conserved during evolution than the amino acid sequences from which they are deduced. In this paper, the hydropathy profiles are used to compare structures of membrane proteins or families of membrane proteins. A technique is proposed that computes the optimal alignment of hydropathy profiles without making use of the underlying sequences. The results show that two membrane proteins with only marginal sequence identity or two non-related families of membrane proteins can have very similar hydropathy profiles, indicating similar global structures. Two parameters are defined that measure differences between hydropathy profiles. The Structure Divergence Score (SDS) provides a measure for the divergence in profiles that reflect one and the same global structure. The SDS is derived from the individual hydropathy profiles of the members of a homologous protein family that are believed to share the same structure. The Profile Difference Score (PDS) quantifies the difference between two hydropathy profiles. Comparison of the PDS of the optimal alignment of the hydropathy profiles of two families of membrane proteins with the SDSs of the two families provides a criterion for structural similarity. Using this technique, pairwise alignment of the family profiles of eight families of secondary transporters suggests that the families fall into four structural classes.

Keywords: membrane protein structure, hydropathy profile, alignment, structural classes, secondary transporters.

Abbreviations: SDS, structure divergence score; PDS, profile difference score.

Introduction

Known structures of many integral membrane proteins show the same basic architecture even though the structures may be quite different; the proteins consist of bundles of membrane spanning hydrophobic α-helices connected by hydrophilic loops. This architecture is not related to the function of the proteins and is observed in independent entities as well as in domains or subunits of multi-component assemblies. The structures of the light driven proton pump bacteriorhodopsin of *Halobacterium salinarium* [1], the photosynthetic reaction centre of purple bacteria that functions as a light driven electron pump [2, 3], the cytochrome *c* oxidases of *Paracoccus denitrificans* and bovine heart mitochondria [4, 5] and the light-harvesting complexes of plant and photosynthetic bacteria [6, 7] show that in the environment of the membrane bundles of transmembrane helices provide a suitable and flexible skeleton to build enzymes with a variety of biological functions associated with the cell membrane. The relatively simple architecture of integral membrane proteins is the reason for the success of secondary structure prediction methods that are based on the hydropathy profile of the primary structure and that were initiated by Kyte and Doolittle [8]. An α-helix that spans the hydrophobic core of the phospholipid bilayer requires a stretch of about 20 residues with a high average hydrophobicity while the inter helical loops that contact the water phase have lower average hydrophobicities. In the hydropathy profile this results in alternating regions of high and low average hydrophobicity which provides a fingerprint of the structure of the protein.

The number of membrane proteins for which the amino acid sequence is known by far exceeds the number of membrane proteins for which a 3-dimensional structure is known. Multiple sequence alignments have classified the proteins in many families that are homologous at the genetic level. It is believed that these homologous proteins are also structurally similar and the few available examples support this view. Bacteriorhodopsin and halorhodopsin, a light driven chloride pump, are homologous proteins and at 7 Å resolution the fold of the two proteins is similar [1, 9]. The L and M subunits of the reaction centres of *Rhodopseudomonas viridis* and *Rhodobacter sphaeroides* are genetically homologous proteins and their 3-dimensional structures are almost super imposable [2, 3] and, finally, the crystal structures of the cytochrome *c* oxidase complex of the bacterium *Paracoccus denitrificans* [4] and its eukaryotic homologue from bovine heart mitochondria [5] show a similar folding of the integral membrane subunits I, II and III. In these examples the sequence identity of the corresponding proteins is only 30–60%. Nevertheless, the hydropathy profiles of these sequences reflect the same global structure and it has been noted before that within a family of membrane proteins the hydropathy profile of the amino acid sequences is much better conserved in the course of evolution than the sequences themselves. As a consequence, the average hydropathy profile of the members of a family, or the family profile, which is obtained from the multiple sequence alignment, will be the best fingerprint of the global structure shared by the proteins in the family. In addition, the set of individual profiles of the members of a family gives an estimate of the divergence in hydropathy profiles that reflect one and the same global structure.

The large tolerance of 3D structure towards changes in amino acid sequence may result in levels of sequence similarities between two proteins from one family that cannot be distinguished from the similarities of random sequences. Alternatively, convergent evolution may have resulted in similar structures of genetically non-related proteins. In both

*To whom correspondence should be addressed.

0968-7688/98 $12.00

cases, the actual similarity of the structures is reflected in the hydropathy profiles of the amino acid sequences. In this paper we explore the possibility to show structural (dis)similarity between membrane proteins by comparing hydropathy profiles. As in comparing amino acid sequences, a fair comparison of hydropathy profiles should take care of inserts and deletions in the polypeptide chains. A technique is proposed to find the optimal alignment of two hydropathy profiles that does not make use of the underlying sequences. The technique will be used to compare averaged hydropathy profiles of families of membrane proteins since we propose that these provide the best fingerprint of the corresponding structures. A quantitative assessment of structural (dis)similarity is sought by comparing the difference between the optimal aligned family profiles and the divergence observed in the profiles of the members in the families.

Results

Hydropathy profile conservation and divergence

Four families of membrane proteins with different functions and with members from different biological origin were selected to show that conservation of hydropathy profiles is a general feature of membrane proteins and to define a measure of divergence of profiles within a family (tables 1, 2). Subunit I (CoxI family) is the central catalytic subunit in the terminal oxidase complexes. The family contains cytochrome *c* oxidases of both prokaryotic and eukaryotic origin and also quinol oxidases of bacteria and archea. The L and M subunits (PufLM family) are two homologous subunits that form the membrane bound core of the reaction centres of phototrophic bacteria. SecY is the major integral membrane subunit of the bacterial preprotein export machinery (SecY family) and, finally, the members of the Glus family are secondary solute transporters. The family contains transporters for glutamate found in bacteria and the central nervous system besides bacterial dicarboxylate transporters.

Members of the families were selected such that pairwise sequence identities were between 20% and 75% (table 1). The distribution of the pairwise sequences shows that most sequences share between 20 and 40% identity (figure 1). The number of residues that is conserved throughout these four families ranges from 15 for the family of the preprotein translocase subunit SecY to 46 for the terminal oxidase subunit I family (table 1).

Figure 2 shows the family hydropathy profiles of the four families (bold lines) together with the hydropathy profiles of the individual members (thin lines). Regions of high and low hydrophobicity coincide in all of the members which results in a family profile that shows the same pattern as each of the members. Clearly, the hydropathy profile is a well conserved property within a family of membrane proteins even though the number of conserved residues in the family is well below 10% and the pairwise sequence identity as low as 20%. The position of the transmembrane helices observed in the crystal structure of the cytochrome *c* oxidase subunits I and the L and M subunits of the bacterial reaction centres correlate nicely with the regions of high average hydrophobicity in the family profiles emphasizing that the family profile reflects the global structure of the family. Gaps in the alignments are almost exclusively found in the regions with the lower hydrophobicity that correspond to the loops that connect the transmembrane segments.

The individual hydropathy profiles in figure 2 give a qualitative impression of the divergence of the hydropathy profiles within a family. The Structure Divergence Score (SDS) quantifies the divergence by averaging the square of the differences between the family profile and the individual profiles at each position. The SDS is computed as follows. First, the hydropathy profile of the family is computed by multiple sequence alignment of the individual sequences and, subsequently, calculation of the average hydrophobicities H_j of the residues in windows j (bold lines in figure 2). Subscript j correlates with the positions in the alignment. Second, the profiles of the individual sequences i are computed using the same window size and aligned with the family profile by

Table 1. Primary sequence and hydropathy profile analysis of families of membrane proteins used in this study.

	Size		Sequence identity		Profile similarity	
Family[a]	Members	Residues	Family (residues)	Pairwise[b] (%)	SDS (hydrophobicity units)	Range (PDS) (hydrophobicity units)
Cyt oxidase subunit I (CoxI)	17	512–788	46	39	0.106[c]	0.092–0.125[c]
Reaction centre subunits L/M (PufLM)	12	274–330	19	31	0.091	0.061–0.117
Protein translocase subunit Y (SecY)	18	409–492	15	35	0.138	0.093–0.187
Glutamate transporters (Glus)	11	415–573	24	27	0.136[c]	0.102–0.159[c]
Citrate/ketoglutarate transporters (CitKgl)	12	425–500	29	28	0.117	0.097–0.170
Galactoside transporters (Gph)	8	463–641	21	25	0.113[c]	0.085–0.154[c]
Sugar transporters (Sugar)	15	464–567	21	26	0.148[c]	0.123–0.172[c]
Tetracycline antiporters (Tetracyc)	10	388–405	36	43	0.101	0.081–0.133
Gluconate transporters (Gluconat)	8	427–454	22	32	0.101	0.081–0.149
Neurotransmitter transporters (Snf)	7	599–632	144	43	0.097	0.084–0.115
Amino acid transporters (AmAc)	30	456–663	10	29	0.126[c]	0.101–0.165[c]

[a]The members of the families are listed in table 2.
[b]Median of the pairwise sequence identity distribution (see figure 1).
[c]Calculated for the part of the sequences common to all members.

introducing the gaps observed in the amino acid sequence in the individual profile. This results in a set of average hydrophobicities h_j^i in which j corresponds to the position in the alignment and i to the individual sequence. Because of the introduction of the gaps a value for h_j^i is not defined for every value of j. The thin lines in the examples in figure 2 represent plots of h_j^i versus j. Then the SDS is defined as

$$\mathrm{SDS} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{W}(H_j - h_j{}^i)^2}{\sum_{i=1}^{n} w_i}} \qquad (h_j{}^i \text{ defined}) \qquad (1)$$

Table 2. Membrane protein families used in this study.[a]

Family	Members (Accession number)	References
CoxI	CTADIIpd(Y07533), COX1bt(P00396), COXArl(Q08855), COXIam(P80440), COXIan(P00402), COXIdy(P00402), COXIhh(P33518), COXIzm(P08742), COXNbj(P98000), CTADbf(Q04440), CTADbs(P24010), CAABps(P16262),CTADspcc(Q06473), CYAAaa(P8009), CYOBec(J05492), QOXBbs(P34956), SOXMsa(P39481)	25
PufLM	PUFLca(P11695), PUFLes(P26280), PUFLrc(K01184), PUFLrhs(P02954), PUFLru(J03731), PUFLrv(P06009), PUFMca(X07847), PUFMes(X57597), PUFMrc(K01184), PUFMrhs(X63405), PUFMru(J03731), PUFMrv(P06010)	
SecY	SECYbc(P38375), SECYbfl(P38376), SECYbs(P16336), SECYcc(P46249), SECYcp(P25014), SECYcph(P28527), SECYct(P28539), SECYec(P03844), SECYhi(P43804), SECYll(P27148), SECYmc(P10250), SECYmg(P47416), SECYml(P33108), SECYpl(P28540), SECYps(P38397), SECYsca(Q05217), SECYscc(P46785), SECYspcc(P31159)	26
Glus	EAAC1oc(P31597), EAAT4hs(P48664), GLASTrn(P31596), GLT1rn(P31596), GLTPbs(P39817), GLTPec(P21345), GLTPse(X92556), GLTTbst(P24943), DCTAec(P37312), DCTArs(P31601), SATThs (P43007)	16, 27
CitKgl	4mpPbce(U29532), CitAec(P07661), CitHkp(P16482), DcaTmt(Z95390), y418hi(P44699), YhjEmt(Z93777), HmtPhi(P44610), KgtPec(P17448), KgtPhp(AE000616), PcaTpsp(U48776), ProPec(P30848), YhjEec(P37643)	
Gph	LACSldb(P22733), LACSll(-), LACSst(PM23009), MELBec(P02921), MELBkp(Q02581), RAFPpp(P43466), XYLPlp(-), XYNCbs(-)	23
Sugar	AraEec(P09830), GalPec(P37021), GLCPssp(P15729), GLFzmo(P21906), GLUTgg(P46896), GLUT1gg(P46896), GLUT2mm(P14246), GLUT3cf(P47842), GTR11d(Q01440), HGT1kl(P49374), Hup1ck(Pp15686), HXT13sc(P39924), HXT3sc(P32466), STP1at(P23586), XylEec(P090908), yfiGbs(P54723)	18
TetraCyc	Bmr1bs(P33449), Bmr2bs(P39843), NorAsta(P21191), TCR1ec(P02982), TCR3ec(P02981), TCR4so(P33733), TCR5ec(Q07282), TCR7va(P51563), TCR8pm(P51564), TetAec(P02980)	
Gluconate	GNTPbs(P12012), GNTPec(P39373), GNTPhi(D64108), GNTTec(P39835), GNTUec(P46858), DSDXec(P08555), YJGTec(P39344). YJHFec(P39357)	17
Snf	GAT-1hs(P30531), GAT-2mm(P31649), GAT-3hs(P48066), BGT-1hs(P48065), DAT1hs(Q01959), NAT1hs(P23975), NTS-1rn(P31652)	28
AmAc	Alp1sc(P38971), AnsPsty(P40812), AroPec(P15993), BAP2sc(P38084), CAN1cal(P43059), CAN1sc(P04817), CycAec(P39312), DIP5sc(P53388), GabPbs(P46349), GabPec(P25527), GAP1sc(P19145), GNP1sc(P48813), HIP1sc(P06775), HutMbs(P42087), INA1th(P34054), LYP1sc(P32487), LysPec(P25737), PAP1sc(P41815), PhePec(P24207), ProYsty(P37460), PUT4sc(P15380), PUTXen(P18696), RocCbs(P39636), RocEbs(P39137), TAT2sc(P38967), VAL1sc(P38085), YBYsc(P38090), YCC5sc(P25376), YFF5sc(P43548), yifKec(P27837)	

[a]Identifiers are followed by a code for the organism as follows: pd, *Paracoccus denitrificans*; bt, *Bos taurus*; rl, *Rhizobium leguminosarum*; am, *Allomyces macrogynus*; an, *Aspergillus nidulans*; dy, *Drospohila yacuba*; hh, *Halobacterium halobium*; zm, *Zea mays*; bj, *Bradyrhizobium japonicum*; bf, *Bacillus firmus*; bs, *Bacillus subtilis*; ps, *Bacterium PS-3*; spcc, *Synechocystis PCC*; aa, *Acetobacter acet*; ec, *Escherichia coli*; sa, *Sulfolobis acidocaldarius*; ca, *Chloroflexus aurantiacus*; es, *Erythrobacter sp*; rc, *Rhodobacter capsulatus*; rhs, *Rhodobacter sphaeroides*; ru, *Rhodospirillum rubrum*; rv, *Rhodopseudomonas viridis*; bc, *Bacillus C-125*; bfl, *Brevibacterium flavum*; cc, *Cyanidium caldarium;* cp, *Cyanophora paradoxa;* cph, *Crytomonas phi;* ct, *Chlamydia trachomatis;* hi, *Heamophilus influenzae;* ll, *Lactococcus lactis;* mc, *mycoplasma capricolum*; mg, *Mycoplasma genitalium*; ml, *Micrococcus luteus*; pl, *Pavlova lutherii*; ps, *Pyrenomonas salina*; sca, *Staphylococcus carnosus*; scc, *Streptomyces coelicolor*; oc, *Oryctolagus cuniculus*; hs, *Homo sapiens*; rn, *Rattus norvegicus*; se, *Saccharopolyspora erythraea*; bst, *Bacillus stearothermophlus*; rs, *Rizobium sp*; bce, *Burkholderia cepacia*; kp, *Klebsiella pneumoniae*; mt, *Mycobacterium tuberculosis*; hp, *Helicobacter pylori*; psp, *Pseudomonas putida, ldb, Lactobacillus delbruecki*; st, *Streptococcus thermophilus*; pp, *Pediococcus pentosaceus*; lp, *Lactobacillus plantarum*; ssp, *Synechocystus sp*; zmo, *Zymomonas mobilis*; mm, *Mus musculus*; gg, *Gallus gallus*; cf, *Canus familiaris*; id, *Leishmania donovani*; kl, *Kluyveromyces lactis*; ck, *Chlorella kessleri*; sc, *Saccharomyces cerevisiae*; at, *Arabidopsis thaliana*; sta, *Staphylococcus aureus*; so, *Salmonella ordonez*; va, *Vibrio anguillarum*; pm, *Pasteurella multocida*; sty, *Salmonella typhimurium*; cal, *Candida albicans*; th, *Trichoderma harzianum*; en, *Emericella nidulans.*

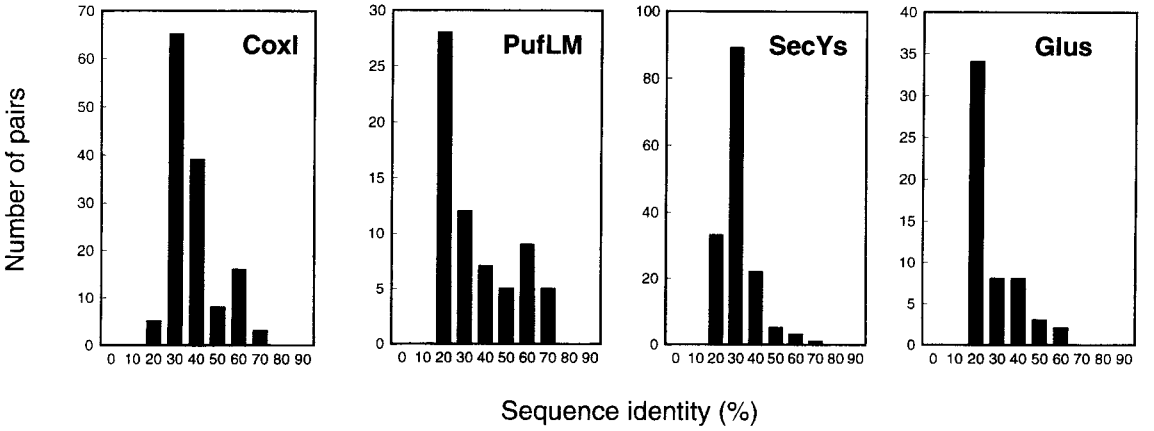

Figure 1. Distribution of the pairwise sequence identity in four families of membrane proteins. A bar labelled '40' indicates the number of pairs in the multiple sequence alignment with pairwise sequence identities between 40 and 50%. The families are defined in Tables 1 and 2.

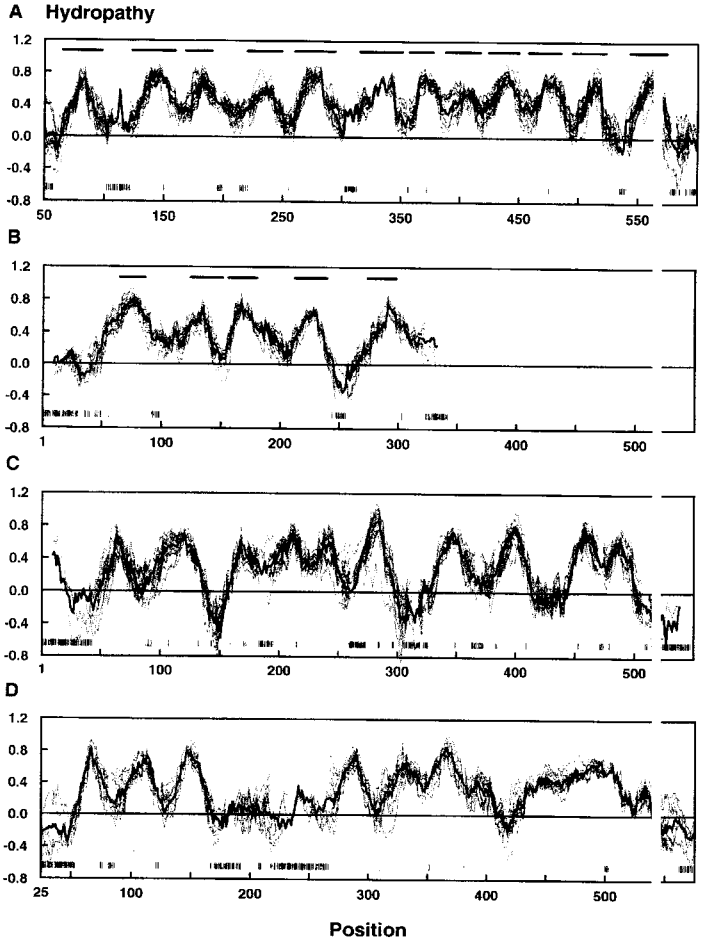

Figure 2. Hydropathy profile analysis of four families of membrane proteins. The family hydropathy profiles of the CoxI (A), PufLM (B), SecY (C) and Glus (D) families defined in Table 2 are given in bold. The thin lines represent the individual profiles of the members. Horizontal bars (top) in A and B indicate the position of the transmembrane α-helices observed in the crystal structure of the cytochrome *c* oxidase subunit I of *Paracoccus denitrificans* (COXIpd) [4] and bovine heart (COXIbt) [5] and the L and M subunits of the reaction centres of *Rhodobacter sphaeroides* (PUFLrsp, PUFMrsp) [3] and *Rhodopseudomonas viridis* (PUFLrv, PUFMrv) [2], respectively. Vertical bars (bottom) indicate positions in the multiple sequence alignment where a gap is found in any of the sequences.

in which n is the number of sequences in the family, W the number of windows in the alignment and w_i the number of windows in sequence i. The square root provides that the unit of the SDS is the unit of the hydrophobicity scale. The structural divergence of the families presented in figure 2 is remarkably similar with SDSs ranging from 0.091 – 0.138 hydrophobicity units (table 1).

Hydropathy profile alignment

Thus far comparisons between hydropathy profiles were restricted to members of a family since the underlying sequence alignment provides a way to align the profiles. To circumvent this restriction, we propose to align hydropathy profiles directly by a similar approach used to align amino acid sequences. Procedures for the alignment of the amino acid sequences of two proteins (or actually nucleotide sequences of two genes) try to mimic an evolutionary pathway by which one sequence is converted into the other while minimizing a cost of conversion [10]. The replacement of one residue for another is costless when the two are the same and is assigned a penalty when different. Furthermore, penalties are given for the introduction of inserts and deletions which results in the gaps in the two sequences. The residue replacements and gaps accumulate in a cost for each possible sequence conversion and the pathway with the lowest cost is taken to be the optimal alignment. A hydropathy profile is a 1-dimensional array of numbers rather than characters, but two profiles can be aligned using the same technique. The cost $c_{i,j}$ associated with the replacement of two residues is replaced by the (absolute) numerical difference in hydropathy h at positions i and j as follows

$$c_{i,j} = | h_i{}^a - h_j{}^b | \qquad (2)$$

in which superscripts a and b reflect the two profiles. Unlike in the case of the alignment of amino acid sequences the replacement cost in hydropathy profile alignment is a continuous function. The costs for the introduction of gaps is related to the hydrophobicity scale. Efficient computational techniques to find the optimal alignment have been described [11, 12, 13].

The quality of the alignment is measured by the Profile Difference Score (PDS) that measures the difference between the two profiles as the average square of the differences at each position in the alignment,

$$\mathrm{PDS} = \sqrt{\frac{\sum_{j=1}^{w} (h_j{}^a - h_j{}^b)^2}{w_a + w_b - W}} \qquad (h_j{}^1,\ h_j{}^2 \text{ defined}) \qquad (3)$$

in which $h_j^{a,b}$ are the hydrophobicity values of the two profiles at position j, $w_{a,b}$ the number of values of $h_j^{a,b}$ and W the number of positions in the alignment. The PDS depends both on the window size used to compute the profiles and on the gap costs. The SDS defined above and the PDS are similar parameters expressing differences between hydropathy plots that can be compared directly. The PDS can be used in general to compare two profiles. The last column of table 1 gives the range of PDSs of the family and the individual profiles that were aligned using the multiple sequence alignment.

To show that the hydropathy profile alignment procedure gives a satisfactory result the profiles of the two members of the SecY family with the lowest pairwise sequence identity were aligned and the result was compared with the profile alignment based on the multiple sequence alignment. It is assumed that the latter is correct because it is based on the pairwise sequence alignment of all the members in the family. The SecY subunits of *Cyanophora paradoxa* SECYcp and *Mycoplasma genitalium* SECYmg share 22% identical residues in the multiple sequence alignment and the hydropathy profiles based upon this alignment are shown in figure 3A. The PDS equals 0.253. The two hydropathy profiles of the primary sequences were aligned by the procedure outlined above using gap costs for opening and extending a gap of 0.8 and 0.4 hydrophobicity units, respectively. Figure 3B shows that the procedure results in virtually the same alignment with a PDS of 0.216. In fact, the hydropathy profile alignment may be the

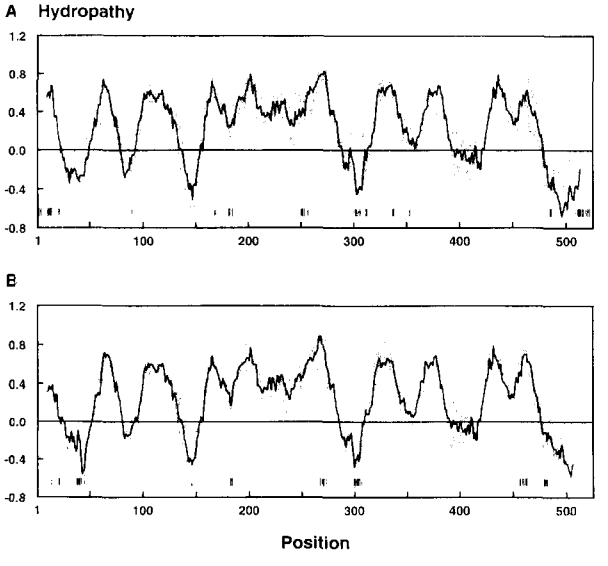

Figure 3. Hydropathy profile alignment of two members of the SecY family. The hydropathy profiles of the protein translocase SecY subunits of *Cyanophora paradoxa* SECYcp and *Mycoplasma genitalium* SECYmg are shown as thin lines and the average of the two as bold lines. The alignments are based on the multiple sequence alignment of the complete SecY family (A) and direct alignment of the two hydropathy profiles as described in the text (B). Vertical bars mark the positions of gaps in any of the two sequences.

Table 3. Optimal alignments of the family hydropathy profiles presented in figure 2[a].

	CoxI	PufLM	SecY	Glus
CoxI	**0.106**	3.20	2.32	3.81
PufLM	0.177	**0.091**	1.21	2.14
SecY	0.188	0.129	**0.138**	2.69
Glus	0.238	0.169	0.225	**0.136**

[a]Indicated in bold are the SDSs of the family hydropathy profiles. The lower off-diagonal elements contain the PDSs of the optimal alignments while the upper off-diagonal elements contain the results of the S-test.

better alignment because it contains less spurious gaps which are a consequence of the 'once a gap, always a gap' principle used in the progressive multiple sequence alignment procedure [14]. The result of the hydropathy profile alignment can be translated back to a sequence alignment. This results in a sequence identity of 16%.

Alignment of family profiles

The alignment of the hydropathy profiles of the two SecYs in figure 3B shows that in spite of the low sequence identity of the proteins the profiles are very similar suggesting a similar structure which is consistent with the fact that these two proteins are homologues. The example suggests that visual inspection of aligned profiles of apparently unrelated families may be a good way to decide on structural similarity. However, visual inspection is prone to subjective variability and we have investigated the possibility to catch the (dis)similarity of family hydropathy profiles in a parameter based on the SDS and PDS parameters defined above to come to a more quantitative, and objective, criterion. The idea is to compare the difference between optimally aligned family profiles (PDS) with the divergence observed in the hydropathy profiles of the members of the families (SDS). The crystal structure of members of the CoxI and PufLM families are different and it is likely that the structure of the Glus and SecY families also represent different structures because of their different functions. Pairwise alignments of the family profiles of the CoxI, PufLM, SecY and Glus families results in PDSs that are significantly higher than the divergence of profiles observed within the families (SDSs) except for the case of the SecY and PufLM families (table 3). The comparison beween PDS and SDS is expressed in a similarity test (*S*-test) which is defined as the ratio of the square of the PDS and the average square of the SDSs of the two families

$$S = \frac{2 * PDS_{a,b}{}^2}{SDS_a{}^2 + SDS_b{}^2} \qquad (4)$$

The values for *S* are two or higher (table 3). Visual inspection of the optimal alignments shows that the family profiles are clearly different. Two examples are shown in figures 4A and B. Transmembrane segments IV around position 240 (figure 4A), VIII (position 450) and XII (position 600) of the CoxI family do not have their counterparts in the Glus family. Similarly, transmembrane segments VIII (position 425) and X (position 490) are not observed in the SecY family. In both alignments

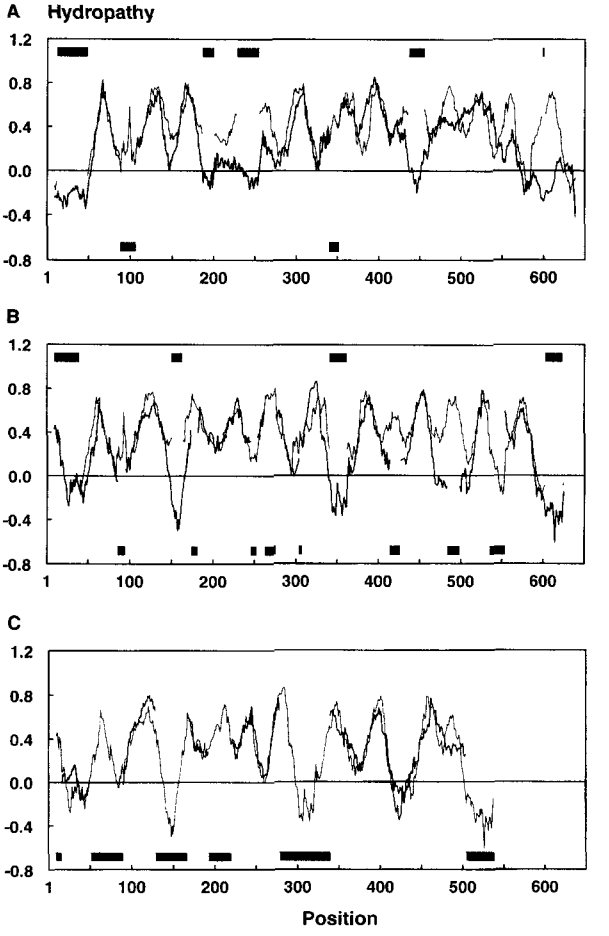

Figure 4. Optimal alignment of the CoxI/Glus (A), CoxI/SecY (B) and SecY/PufLM (C) family hydropathy profiles. The profiles are indicated as bold and thin lines, respectively. The bars at the top and bottom indicate the gaps in the two profiles.

Table 4. Optimal alignments of the family hydropathy profles of secondary transporter families.[a]

	Glus	CitKgl	Gph	Sugar	Tetracyc	Gluconat	SNF	AmAc
Glus	**0.136**	2.92	3.23	1.98	4.65	3.65	3.31	4.03
CitKgl	0.217	**0.117**	1.07	1.00	1.09	4.28	2.45	2.18
Gph	0.225	0.119	**0.113**	0.92	0.98	3.51	2.57	2.33
Sugar	0.200	0.133	0.126	**0.148**	0.76	2.83	2.77	3.18
Tetracyc	0.258	0.114	0.106	0.110	**0.101**	3.80	2.95	1.87
Gluconat	0.229	0.226	0.201	0.213	0.197	**0.101**	4.32	1.95
SNF	0.215	0.168	0.169	0.208	0.170	0.206	**0.097**	1.28
AmAc	0.263	0.180	0.183	0.245	0.156	0.159	0.127	**0.126**

[a]Indicated in bold are the SDSs of the family hydropathy profiles. The lower off-diagonal elements contain the PDSs of the optimal alignments of the corresponding family profiles while the upper off-diagonal elements contain the results of the S-test. Families that potentially belong to the same structural class were grouped in boxes.

gaps are found in hydrophobic peaks which is in contrast to what is observed in the hydropathy profiles of the members of homologous families (see figure 2). Overall, the optimal alignments still reveal distinct differences between the profiles. The optimal alignment of the SecY and PufLM families shows that the shorter PufLM profile is scattered over the longer SecY profile (figure 4C). Such an alignment is meaningless in terms of structural similarity and the low PDS and value of the *S*-test are a consequence of the multitude of possibilities by which the five hydrophobic regions in the PufLM profile can be distributed over the 10 similar regions in the SecY profile. This number of possibilities amounts to 252 (10!/(10-5)!/5!). In conclusion, the different structures of the four families are reflected in different hydropathy profiles which, after optimal alignment of the profiles, result in PDSs that are significantly larger than the corresponding SDSs. Artefacts may occur when the structure of two families contains very different numbers of transmembrane segments.

Secondary transporters facilitate the translocation of solutes across biological membranes (for a review, [15]). The genes coding for many secondary transporters have been described and sequence homology has classified them in many different families. No structure of any of these transporters is known to date but, because of similar function, the families may have much more similar structures than the CoxI, PufLM, SecY and Glus families compared above. The eight families of secondary transporters listed in table 1 contain at least seven members with pairwise sequence identities between 20 and 75%. The membrane embedded part of the transporter molecules is more or less of the same length consisting of about 450 residues. Therefore, it is expected that the number of transmembrane segments is not very different (see Discussion). The SDSs of the families are similar to those of the families discussed above (table 1). Table 4 presents the analysis of the pairwise alignments of the family profiles. The CitKgl, Gph, Sugar and Tetracyc families form a cluster in

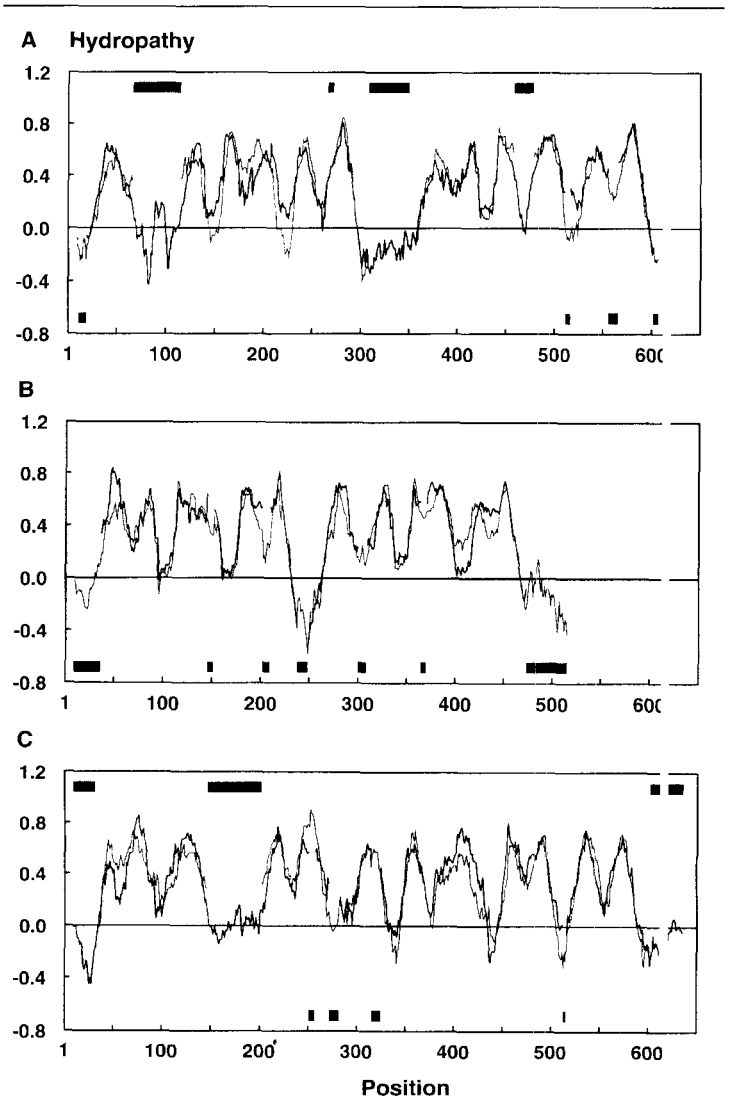

Figure 5. Optimal alignment of family hydropathy profiles of secondary transporter families. (A) Gph and Sugar (bold). (B) CitKgl and Tetracyc (bold). (C) AmAc and Snf (bold). The bars indicate the gaps in the profiles.

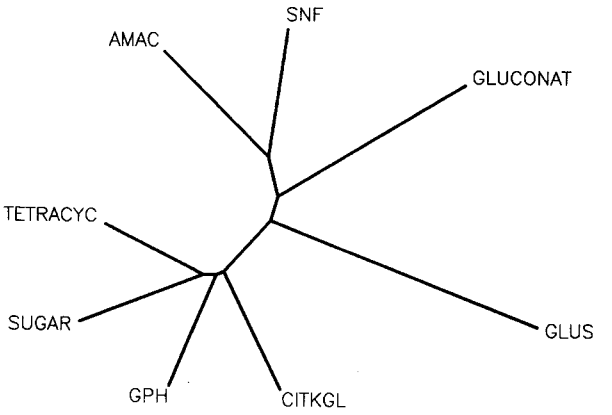

Figure 6. Cluster analysis of the family profiles of eight secondary transporter families. The PDS matrix in Table 4 was used as the input for the Neighbor program and the tree was plotted by the Drawtree program. Both programs are part of the Phylip package [29].

which the PDSs are in the same range as the SDSs of the family profiles. The similarity tests result in values close to 1. The alignments of the Gph and Sugar families and of the CitKgl and Tetracyc families are shown in figures 5A and 5B, respectively. The hydrophobic regions in the two profiles nicely overlap and the pattern of the peak heights is almost identical. The gaps are mostly found in the hydrophilic regions that in the structure correspond to the loops that connect the transmembrane segments. Overall the profiles are very similar, strongly suggesting similar structures. Importantly, none of the alignments of the four families in this cluster with the other four families results in a similarly low PDS, indicating that this cluster represent a separate structural class that is different from the structures of the other families. One other pair of families results in a PDS that is in the same range as the SDSs even though the *S*-test results in a slightly higher value. The alignment of the Snf and AmAc families is shown in figure 5C and the high similarity suggests that also these two families represent a separate structural class. Alignment of the Glus and Gluconat families with the other families results in values for the *S*-test of at least two suggesting that each of these families represents a separate structural class. Visual inspection of the pairwise alignments of the transporter families that result in values for the similarity test of two or higher shows that this correlates with family profiles that are clearly different (not shown). The analysis of the eight family profiles is summarized by the cluster analysis presented in figure 6. The four structural classes are easily recognized and it follows that the Glus family is most distant from the other families.

Discussion

The hydropathy profile of a membrane protein makes a link between the amino acid sequence of the polypeptide chain and its 3-dimensional folding. The profile provides a fingerprint of the structure of the protein that may contain information like the number of transmembrane segments, the length of the transmembrane segments, the folding of the secondary structure elements and polarity properties of the secondary structure elements. Clearly, the structure at this level of detail cannot be deduced from the profile and structure prediction based on hydropathy properties of the amino acid sequence is restricted to the membrane topology and amphipathy of transmembrane segments. However, in the reverse order, a specific structure uniquely defines the hydropathy profile which can be used to discriminate between different structures or to show structural similarities between apparent non-homologous proteins or protein families. A fair comparison of two hydropathy profiles should take into account that deletions and insertions may occur. Therefore, a procedure was developed to find the optimal alignment of two hydropathy profiles that is based on similar procedures described for the alignment of two amino acid or nucleotide sequences [10]. The efficiency of the procedure is demonstrated by the alignments shown in figure 3B for two individual profiles and in figure 5 for family profiles. These examples demonstrate that membrane proteins or protein families that at the amino acid sequence level are not or only marginally homologous can have remarkably similar hydropathy profiles, indicative of similar structures.

The hydropathy profile alignment technique was used to compare averaged hydropathy profiles of families of homologous membrane proteins. Homologous proteins are believed to have the same global structure and, consequently, a family of proteins is characterized by a family hydropathy profile which reflects the secondary and tertiary structure elements that form the skeleton that is common to the members of the family. The skeleton is largely insensitive to differences in the amino acid sequence that may give individual members their specific functional details like substrate specificity or, simply, may result from genetic drift during evolution. The differences in primary structure may result in minor distortions of the skeleton or represent the structure at the amino acid side chain resolution. Anyhow, the family hydropathy profile will give a more accurate fingerprint of the global structure than the individual profiles. The family profile may be obtained by averaging the individual hydropathy profiles of the members which removes the noise introduced by the individual amino acid sequences. Reconstruction of the family profile in this way relies heavily on the correct alignment of the sequences, the composition of the family and the number of members in the family. A lower limit of pairwise sequence identity of 20% was used to prevent serious errors in the alignments and, at the upper limit, a cut-off of about 75% sequence identity was used to prevent a bias to almost identical sequences (figure 1). Building the families by adding the members in random order showed that between 5 and 10 sequences result in a family profile that is more or less independent of the composition of the family (not shown). In constructing the families used in this study a lower limit of seven members with pairwise sequence identities within the limits given above was used.

The averaging procedure for obtaining the family profile provides a measure for the difference in individual profiles that still reflect the same global structure. The SDS is a measure of the distance of the member profiles to the family profile. We

have attempted to use this measure as a criterion for structural similarity between two different families of proteins by comparing the SDSs of the two families with the PDS, a similar measure of the difference between the optimal aligned hydropathy profiles of the two families. In this approach, one family profile is treated as a member of the other family and *vice versa* (note that the nominator in equation 3 equals the second summation in the nominator of equation 1). We have correlated the value of the similarity test in which the PDS and SDSs are compared (equation 4) with the visual inspection of the alignment of two family profiles and found that similar profiles result in values for S of about 1, while different profiles result in values of about 2 or higher. The example alignments given in figures 3B, 4 and 5 give a good feel for the features of similar and dissimilar profiles. In similar profiles, the pattern of the peaks is the same and most of the gaps are in the hydrophilic regions. In contrast, dissimilar profiles are characterized by different peak heights, by the absence of hydrophobic regions in one of the two profiles and by large gaps in hydrophobic regions. The observed correlation may not hold when the profiles of two families with very different numbers of transmembrane segments are compared. Occasionally, this may result in a value for S that is too low (e.g. the PufLM and SecY families). To a first approximation this artefact may be circumvented by examining families of which the membrane bound part contains approximately the same number of residues.

The PDS and, consequently, the outcome of the similarity test, depend on the cost parameters for the introduction of gaps used in the alignment procedure. Low costs tend to lower the PDS but may result in the introduction of too many gaps and fragmentation of the profiles, while high costs increase the PDS by inhibiting the introduction of gaps and the alignment procedure. In this study, the cost parameters for opening and extending gaps were selected to discriminate optimally between similar and dissimilar family profiles in the similarity test. Typically, the alignment of similar profiles is much less sensitive to the gap costs than the alignment of dissimilar profiles. Consequently, relatively high costs give the best results in the similarity test. For other purposes, for instance, to find the optimal alignment of the amino acid sequences of two families with similar profiles it may be useful to reduce the gap costs in the alignment procedure.

The paradigm that secondary transporters like the ones presented in this study consist of 12 transmembrane segments is supported by many topological studies of individual transporters from different homologous families. The present analysis of the family hydropathy profiles suggests that within this superfamily classes can be discriminated that have different structures. The different structures may still correlate with a different number of transmembrane segments. The Glus family which is the most distant from the other families (see figure 6) was recently reported to contain 10 transmembrane segments [16] and secondary structure prediction of the Gluconat family, also not related to any of the other families, revealed 12 – 14 transmembrane segments [17]. It is likely that structures consisting of different numbers of transmembrane segments have different tertiary structures, which is in agreement with the results of the similarity tests (table 4). Members of the other six families are all predicted to be 12 helix bundles and for many transporters this is supported by biochemical evidence. The present analysis discriminates two structural classes containing the CitKgl, Gph, Sugar and Tetracyc families and the Snf and AmAc families. In spite of the lack of significant sequence similarity, structural similarity of the families in the former class has been suggested before because of the presence of (i) 12 putative transmembrane spans, (ii) a long central hydrophilic cytoplasmic loop and (iii) a 6 residue long sequence motif that is predicted to be a β-turn [18]. Within one structural class the difference between the families is in the length of the loops that connect the transmembrane segments, e.g. members of the Sugar family contain a pronounced loop around position 475 that is absent in the Gph family (figure 5A). Similarly, the Snf family contains a 50 residue long loop around position 180 that is absent in the AmAc family (figure 5C). In this respect, the hydropathy profile alignments of families belonging to the same structural class may be helpful in resolving the profile into the different transmembrane segments. For instance, the hydrophilic dip around position 475 in the Sugar family profile (figure 5A) indicates that the hydrophobic region around this position in the Gph family profile represents two transmembrane segments.

Experimental procedures

Families of homologous membrane proteins

Homologous proteins were taken from the literature (see Table 2 for references) or by screening the available databases using the Blast facility [19]. The amino acid sequences of a family were aligned with the Clustal W program using the default settings [20]. Sequences were removed from the alignment until the pairwise sequence identities between all members were between 20 and 75%. Pairwise sequence identity is defined as the number of identical amino acid residues between two sequences in the alignment divided by the number of residues of the shortest of the two. Minor manual adjustments to the alignments were made in the N- and C-terminal flanks when unrealistic gap sequences were introduced. These gaps were removed. The N- and C-terminal hydrophilic regions of the members of the secondary transporter families Glus, AmAc and Sugar are very variable in length and average hydrophobicity. These flanking regions were removed. Some members of the terminal oxidases family (CoxI) contain transmembrane segments at the N- and C-terminal flanks in addition to the twelve transmembrane segments common to all members in the family [21, 22]. The family was aligned using the complete sequences after which these flanking regions were removed from the multiple sequence alignment. All single sequences in the family were truncated at identical positions for further use. Similarly, some members of the secondary transporter family Gph contain a C-terminal soluble domain [23]. This family was treated in the same way. The families and their members used in this study are listed in Table 2.

Hydropathy profiles

The normalized consensus hydrophobicity scale of Eisenberg [24] was used. The hydropathy profile of a single sequence is computed by calculating the average hydrophobicity of a continuous stretch of residues (the window), plotting the value at the position of the center residue and, subsequently, sliding the window over the sequence while repeating the procedure. The hydropathy profile of a family of homologous proteins is obtained in a similar way by sliding the window over the positions in the aligned set of sequences and averaging the hydrophobicity of the residues in the window in all sequences while ignoring gaps. A window of 19 positions or residues was used.

To compare the hydropathy profiles of the family and the individual sequences, the profile of the latter was plotted by the following procedure. The hydropathy profile of the member was computed as

described above and, subsequently, the gaps observed in the individual sequence in the multiple sequence alignment were introduced at the corresponding positions in the hydropathy profile plot. The multiple sequence alignment was also used to compare the hydropathy profiles of two members of a family. The two sequences were taken out of the multiple sequence alignment and gaps occurring at the same positions in both sequences were removed. The remaining gaps in each sequence were introduced in the corresponding profile. All these manipulations were done using MemGen, a program made by jl in the Delphi environment (Borland International, Inc, Scotts Valley, USA).

Hydropathy profile alignment

Optimal alignment of profiles is sought by minimizing a cost parameter for the conversion of one profile into the other in a similar way as is done for the alignment of two amino acid or nucleotide sequences [10]. Allowed operations are the replacement of a value of one profile by a value of the other profile and the introduction of gaps. The cost associated with the replacements is taken as the absolute numerical difference between the two hydrophobicity values. The cost for introducing a new gap and extending a gap were 0.8 and 0.4 hydrophobicity units. The minimal cost of conversion and the corresponding alignment were computed by the procedure described by Meyers and Miller [11] which is based on algorithms described by Gotoh [12] and Hirschberg [13]. Alignment procedures were implemented in the Pascal language in the program Profile made by jl in the Delphi environment (Borland International, Inc, Scotts Valley, USA).

Acknowledgements

The authors would like to thank Chris van der Does for the SecY family alignment and Bert Poolman for the Gph family alignment.

This work was supported by the Ministry of Economic Affairs, the Ministry of Education, Culture and Sciences, and the Ministry of Agriculture, Nature Management and Fishery in the framework of an industrial relevant research program of the Netherlands Association of Biotechnology Centres in The Netherlands (ABON) by a grant to D-J. S.

References

1. Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckman, E. and Downing, K. H. (1990) Model for the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. *Journal of Molecular Biology*, **213**, 899–929.
2. Deisenhofer, J., Epp, O., Miki, K., Huber, R. and Michel, H. (1985) Structure of the protein subunits in the photosynthetic reaction center of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature*, **318**, 618–624.
3. Allen, J. P., Feher, G., Yeates, T. O., Komiya, H. and Rees, D. C. (1987) Structure of the reaction center from *Rhodobacter sphaeroides* R-26: the protein subunits. *Proceedings of the National Academy of Sciences, USA*, **84**, 6162–6166.
4. Iwata, S., Ostermeier, C., Ludwig, B. and Michel, H. (1995) Structure at 2.8 Å resolution of cytochrom *c* oxidase of *Paracoccus denitrificans*. *Nature*, **376**, 660–669.
5. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrom c oxidase at 2.8 Å. *Science*, **272**, 1136–1140.
6. Kuhlbrandt, W., Wang, D. N. and Fujiyoshi, Y. (1994) Atomic model of plant light-harvesting complex by electron crystallography. *Nature*, **367**, 614–621.
7. McDermott, G., Prince, S. M., Freer, A. A., Hawthornthwaite-Lawless, A. M., Papiz, M. Z., Cogdell, R. J., Isaacs, N. W. (1995) Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature*, **374**, 517–521.
8. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**, 105–132.
9. Havelka, W. A., Henderson, R. and Oesterhelt, D. (1995) Three-dimensional structure of halorhodopsin at 7 Å resolution. *Journal of Molecular Biology*, **247**, 726–738.
10. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.
11. Myers, E. W. and Miller, W. (1988) Optimal alignments in linear space. *Computer Applications in the Biosciences*, **4**, 11–17.
12. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**, 705–708.
13. Hirschberg, D. S. (1975) A linear space algorithm for computing longest common subsequences. *Communications of the Association of Computing Machinery*, **18**, 341–343.
14. Feng, D-F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, **25**, 351–360.
15. Poolman, B. and Konings, W. N. (1993) Secondary solute transport in bacteria. *Biochimica et Biophysica Acta*, **1183**, 5–39.
16. Slotboom, D-J., Lolkema, J. S. and Konings, W. N. (1996) Membrane topology of the C-terminal half of the neuronal, glial and bacterial glutamate transporter family. *Journal of Biological Chemistry*, **271**, 31317–321321.
17. Peekhaus N., Tong, S., Reizer, J., Saier, M. H., Murray, E. and Conway, T. (1997) Characterization of a novel transporter family that includes multiple *Escherichia coli* gluconate transporters and their homologues. *FEMS Microbiology Letters*, **147**, 233–238.
18. Henderson, P. J. F. (1991) Sugar transport proteins. *Current Opinion in Structural Biology* **1**, 590–601.
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment research tool. *Journal of Molecular Biology*, **215**, 403–410.
20. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Research* **22**, 4673–4680.
21. Chepuri, V. and Gennis, R. B. (1990) The use of gene fusions to determine the topology of all of the subunits of the cytochrome *o* terminal oxidase complex of *Echerichia coli*. *Journal of Biological Chemistry*, **265**, 12978–12986.
22. Lübben, M., Arnaud, S., Castresana, J., Warne, A., Albracht, S. P. J., Saraste, M. (1994) A second terminal oxidase in *Sulfolobus acidocaldarius*. *European Journal of Biochemistry*, **224**, 151–159.
23. Poolman, B., Knol, J., van der Does, C., Henderson, J. F., Liang, W-J., Leblanc, G., Pourcher, T. and Mus-Veteau, I. (1996) Cation and sugar selectivity determinants in a novel family of transport proteins. *Molecular Microbiology*, **19**, 911–922.
24. Eisenberg, D. (1984) Three-dimensional structure of membrane and surface proteins. *Annual Reviews of Biochemistry*, **53**, 595–623.
25. Castresana, J., Lübben, M., Saraste, M. and Higgins, D. G. (1994) Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. *EMBO Journal*, **13**, 2516–2525.
26. Rensing, S. A. and Maier, U. G. (1994) The SecY protein family: comparative analysis and phylogenetic relationships. *Molecular Phylogenetic Evolution*, **3**, 187–191.
27. Kanai, Y., Smith, C. P. and Hediger, M. A. (1994) A new family of neurotransmitter transporter: the high affinity glutamate transporters. *FASEB Journal* **8**, 1450–1459.
28. Reizer, J., Reizer, A. and Saier, M. H. Jr (1994) A functional superfamily of sodium/solute symporters. *Biochimica et Biophysica Acta*, **1197**, 133–166.
29. Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Received 11 July 1997, and in revised form 28 November 1997